Dean&Francis
ISSN 2959-6157

A Comprehensive Investigation of Attention-Based CNNs for Galaxy Morphology Classification

Chenyu Zhao

School of Management, University College London, the United Kingdom
chenyu.zhao.24@ucl.ac.uk

Abstract:

Galaxy morphology classification is a fundamental yet challenging task in astronomical research, as it plays a crucial role in understanding the universe's structure and evolution. The complexity and diversity of galaxy shapes, combined with the large volume of astronomical data, necessitate efficient and accurate automated methods for classification. This paper provides a comprehensive review of attention-based convolutional neural networks (CNNs) for galaxy classification, highlighting their potential to overcome limitations of conventional approaches. This paper systematically surveys three advanced methods: attention-gating, multi-branch attention networks, and dynamic multiscale attention networks, detailing their architectures, mechanisms, and performance gains. These methods enhance feature focus, multi-scale fusion, and interpretability while reducing computational costs. The analysis confirms that attention-based CNNs significantly improve classification accuracy and robustness, making them highly valuable for large-scale astronomical surveys. This review offers valuable insights into current advancements and future directions, contributing to the development of more reliable and efficient galaxy classification systems.

Keywords: Attention CNN, attention-gating, multi-branch attention networks, dynamic multiscale attention network.

1. Introduction

In astrophysics, a galaxy is a complex dynamical system composed of stars, interstellar gas, dust and dark matter bound by gravity, and it is the fundamental building block of the cosmic structure. Current research indicates that there are approximately two trillion observable galaxies in the universe [1].

Galaxy classification is one of the fundamental tasks in astronomical research, but for a long time in the past, astrophysicists mainly relied on manual classification of galaxies, which was not only inefficient but

ISSN 2959-6157

also had a high error rate. Research shows that the team led by Swiss astronomer Fritz Zwicky spent seven years classifying only about 30,000 galaxies, averaging less than 15 per day, with a missed detection rate of over 40% [2]. With the development of technology, to improve the efficiency and accuracy of galaxy classification, researchers began to gradually use neural network technology to classify galaxies from 1990.

In 1992, Storrie-Lombardi et al. first introduced Artificial Neural Networks (ANNs) into galaxy classification. By constructing a three-layer feedforward network, with non-parametric morphological indicators of galaxy images as input parameters and Hubble types as output [3]. In 2015, Dieleman et al. applied the Orientation-Invariant Convolutional Neural Network (CNN) technology to galaxy classification. By inputting Sloan Digital Sky Survey (SDSS) galaxy thumbnails and employing a data augmentation strategy to learn rotational invariance, they constructed a network with four convolutional layers and two fully connected layers, which output 11 fine morphological categories of the Galaxy Zoo [4]. In 2022, Dai et al. classified galaxies using an improved deep residual network (ResNet-26) technique. By reducing the depth of ResNet to 26 layers, increasing the channel width, and introducing a multi-scale feature fusion module, the sensitivity to small-scale structures was enhanced [5]. In 2024, Jiang et al. utilized the Convolutional Vision Transformer (CvT) technology to classify galaxies. By integrating the local feature extraction capability of convolutional neural networks with the global attention mechanism of Transformers, they achieved a collaborative modeling of the multi-scale structures of galaxies [6].

Although many studies in the past have proposed different models or classification methods, there are still many limitations in using conventional CNN methods for galaxy classification, which is mainly stem from the diversity of galaxies and the complexity of images. Astronomical images often contain noise, occlusions, and overlapping celestial bodies, further increasing the difficulty of classification. Attention-based methods have shown significant advantages in galaxy classification. They can adaptively focus on key regions in the image and establish long-distance pixel associations, which is particularly important for identifying the fine structural features of galaxies.

The aim of this paper is to discuss the method of classifying galaxies using attention-based CNN. The remaining part of this article consists of Method, Discussion and Conclusion. In Method section, the author will focus on elaborating the workflow, innovation points and framework of Attention-based CNN. In Discussion section, the current challenges and future prospect in this field will be analyzed, and a summary will be made in Conclusion section.

2. Attention-based CNN Methods for Classifying Galaxies

The latest research indicates that, compared with conventional CNN, attention-based models can significantly enhance the accuracy and interpretability in galaxy classification. Next, the author will introduce three commonly used attention CNN methods in the field of galaxy classification, namely attention-gating, multi-branch attention networks, and dynamic multiscale attention network.

2.1 Attention-gating

In October 2020, Bowles et al. proposed the use of the attention-gating method for galaxy classification [7]. Attention-gating is a neural network that combines the attention mechanism and gating units to enhance the model's focus on and control over key information. The attention mechanism can simulate the focusing ability of human vision, enabling the model to automatically identify and focus on important features and regions in the input data while ignoring irrelevant information [8]. Gating units, on the other hand, can control the flow of information, determining which information should be retained or discarded, and then dynamically adjust the weights of features, further focusing on key features through the attention mechanism [9].

When classifying galaxies, the Attention-gating CNN influences the output methods of each model through range normalization, fine-tuning aggregation, and three attention gating mechanisms [7]. The range normalization and fine-tuning aggregation methods can provide significantly improved attention maps, and the three attention gates can enhance the interpretability of the generated attention maps [10].

Bowles et al. used the image data from the VLA FIRST survey as the dataset. After preprocessing and data augmentation, they compared the performance of the trained Attention-gating CNN architecture with the classical CNN model trained and evaluated by Tang et al. [11].

The experimental results in this paper demonstrated when trained with the MiraBest dataset [12], the performance of the models has all improved, and the number of parameters used by the Attention-gating CNN is less than half that of the classical CNN.

Compared with conventional CNN, this method has the advantages of high parameter efficiency, strong interpretability and strong anti-overfitting ability, which enhances the credibility and debuggability of the model. However, the interpretability of its attention map is greatly affected

by hyperparameters, and different normalization and aggregation methods will significantly affect the clarity and interpretability of the attention map.

2.2 Multi-branch Attention Networks

In January 2021, Zhang et al. proposed the use of a multi-branch attention network for classifying galaxy clusters [13]. This method combines attention and bivariate Gaussian distribution, and classifies the given galaxies by using spatial attention. The multi-branch attention network is divided into a primary branch and an auxiliary branch, both of which adopt ResNet-18 as the backbone network [14].

Researchers used the ResNet architecture in the primary branch to extract advanced feature maps, and represented the output of the last residual block before global average pooling as M to generate the attention map [15]. Eventually, the feature vector extracted by the pooling layer was denoted as V1. In the auxiliary branch, the output of the last global average pooling layer was recorded as V2. V1 and V2 were concatenated to form a 1x1x1024 tensor map, which was then passed to the fully connected layer to predict the classification distribution of three galaxy cluster categories. This study formulated a parameter function using the bivariate Gaussian distribution to guide the model to focus on the central region of the image during the training process.

This study employs a combined loss function of classification and regression to train the model. The overall loss function is expressed as L = xxx, where x, x, and x represent the classification loss, namely the primary loss, auxiliary loss, and fusion loss, respectively, which are used to identify the core type of galaxy clusters in the input X-ray images. The last X represents the regression loss.

This study trained the model using a combined loss function of classification and regression, and the overall loss function is expressed as $L = \alpha_P L_P + \alpha_\alpha L_\alpha + \alpha_f L_f + \alpha_r L_r$ among them, $\alpha_P L_P$, $\alpha_\alpha L_\alpha$, and $\alpha_f L_f$ are classification losses, namely primary loss, auxiliary loss, and fusion loss, respectively, used to identify the core type of galaxy clusters in the input X-ray images. The last one, $\alpha_r L_r$, is the regression loss, which utilizes Cramér distance to incorporate the ordinal relationship among cluster types [16].

Zhang et al. utilized the TNG300 from the IllustrisTNG project as the dataset [17], preprocessed it, and optimized it with Adam. They compared the experimental results with ResNet-18 as the baseline method [13].

Compared with the conventional CNN, this method has the advantages of accurately locating key regions, a dual-branch collaborative mechanism, and a loss function design that integrates domain knowledge. It can suppress the interference of irrelevant background noise and enhance the model's ability to perceive subtle differences.

2.3 Dynamic Multiscale Attention Network

In July 2025, Ma et al. proposed the use of Dynamic Multiscale Attention Network (DMAGNet) for classifying galaxy morphologies [18]. This method combines the Dynamic Large Kernel (DLK), Multiscale Feed-Forward Network (MS-FFN), and Attentional Feature Fusion (AFF) modules to enhance the model's ability to extract both global and local features from galaxy images. The DLK module expands the receptive field by adaptively selecting large convolutional kernels to extract features, the MS-FFN processes multi-scale features, and the AFF module replaces the addition operation in traditional residual connections to achieve more effective feature fusion.

When classifying galaxies, DMAGNet takes 256×256-pixel galaxy images as input, undergoes initial convolution and normalization through the Stem module, and then extracts multi-level features through multiple DMA modules (including DLK, MS-FFN, and AFF). Finally, it outputs the probability distribution of six types of galaxy morphologies through the attention pooling layer and fully connected layer in the Head module [19]. This network significantly improves classification performance while maintaining a low parameter count and computational complexity.

Ma et al. utilized 15,266 galaxy images from the Galaxy Zoo DECaLS project as the dataset [20], which was classified into six categories: Edge-On, Cigar, In-Between, Round, Spiral, and Merger. To address the issue of class imbalance, data augmentation operations such as horizontal flipping, vertical flipping, and random rotation were performed on the minority classes, ultimately constructing a balanced dataset containing 23,610 images. They compared DMAGNet with various mainstream networks including ResNet, MaxVit, and TransNext on the same test set.

The experimental results show that DMAGNet achieves an accuracy of 97.1%, a recall rate of 96.8%, and an F1 score of 96.8% on the test set, significantly outperforming other comparison models. Particularly in terms of the number of parameters (9.4M) and computational cost (4.8 GFLOPs), DMAGNet demonstrates high efficiency while maintaining high performance. The ablation experiments further verify that the AFF module contributes the most to the improvement in accuracy, while the DLK module can significantly reduce the model complexity while still maintaining high classification performance.

Compared with the conventional CNN, this method has a stronger ability to fuse multi-scale features, higher classi-

ISSN 2959-6157

fication accuracy and better computational efficiency. Its attention mechanism and dynamic convolutional kernel design enhance the model's perception ability of the details and global structure of galaxy morphology. Meanwhile, it shows better robustness when facing category imbalance and image quality differences. However, the model still has some confusion when dealing with galaxies with highly similar morphologies (such as Cigar and Edge-On). In the future, it can be further optimized by introducing more fine-grained feature representations or fusing multi-source data.

3. Discussion

3.1 Current Limitations of Attention-based CNN

Based on the above analysis of the attention-based CNN, it can be noted that although attention-gating, multi-branch attention networks, and large kernel attention have significantly improved the efficiency, accuracy, and robustness of galaxy classification, they still face some limitations. In this section, the authors will analyze these limitations and provide relevant suggestions for future improvements.

3.1.1 The Unstable Interpretability of Attention Maps

The interpretability of attention maps is highly dependent on the choice of normalization functions and aggregation methods, although these choices have relatively little impact on classification performance metrics. In essence, this constitutes a paradox of the attention mechanism in practical applications.

Different normalization methods determine the distribution characteristics of attention weights. Softmax generates sparse and sharp attention maps, which, although they may enhance the discriminative power of key features, mask secondary features such as weak jets or extended petal-like structures, significantly reducing the richness of interpretation; the Sigmoid function, due to its saturation property, often produces fuzzy and diffuse attention responses, making it difficult to clearly define the precise boundaries of the regions of interest for the model; while Range normalization retains the relative gradient of feature responses, generating smoother and more continuous attention maps, thus aligning more closely with the visual judgment process of human experts. In 2019, Schlemper et al. abandoned the use of softmax as the normalization method for attention gating due to the sparsity of the attention maps caused by the Softmax normalization method.

The aggregation method indirectly affects the spatial focus degree and hierarchical consistency of the attention map by changing the fusion mode of the outputs of different-level attention gates. This leads to a practical dilemma: in order to obtain an easily understandable attention map, researchers have to give up the configuration that is slightly better in quantitative indicators. As a result, the model debugging process not only needs to optimize performance but also "debug" its interpretative behavior, increasing the complexity of usage.

Therefore, the selection of hyperparameters essentially involves a trade-off between model performance and interpretability, which undermines the credibility of the attention mechanism as a stable and interpretable tool, highlighting the necessity of developing attention architectures that are more robust to hyperparameters or can adaptively generate the most interpretable results.

3.1.2 High Computational Costs and Structural Complexity

Although multi-branch has achieved performance improvements, it comes with high computational costs and structural complexity. Firstly, its dual-branch design essentially requires the parallel operation of two independent ResNet-18 backbone networks: the primary branch processes the global image, while the auxiliary branch handles the core region cropped by attention and Gaussian masks. This means that the model's parameter count and the computational load during forward propagation nearly double, resulting in significantly longer training and inference times compared to the single-branch baseline model.

Secondly, the generation and fusion process of the attention map and Gaussian mask further increase the computational layers and complexity. Generating the class activation map requires weighted summation and upsampling of the final feature map, while generating the Gaussian mask involves calculating the probability distribution of each pixel in the image relative to the brightest point. All these operations require additional computing resources.

More crucially, this architecture introduces multiple hyperparameters that require fine-tuning, such as τ_1 which determines the binarization threshold of the attention map, λ which controls the range of the Gaussian distribution, and τ_2 which decides the final mask. These hyperparameters are interdependent, and their optimal values are hard to determine directly, necessitating extensive trial-and-error experiments to identify them, which significantly increases the cost of model development and tuning.

Ultimately, this complexity directly affects the stability of training. The joint training of multiple components such as the main branch, auxiliary branch, and attention module makes the optimization landscape of the loss function more complex, with longer gradient flow paths and

a greater tendency to fluctuate, which may lead to slow convergence or getting stuck in local optima during the training process.

3.2 Future Prospects of Attention-based CNN

3.2.1 Improving the Stability and Reliability of the Attention Mechanism Through Multi-Dimensional Strategies

Firstly, an adaptive normalization mechanism can be developed. By designing a learnable normalization module, the model can dynamically select or fuse different normalization methods based on the features of the input image, thus avoiding reliance on manual selection. At the same time, the clarity of the attention map is introduced as an auxiliary loss term to enhance the visual interpretability of the attention map while optimizing the classification accuracy.

Secondly, visual-semantic alignment supervision can be introduced. By using the key regions annotated by experts as weak supervision signals or by adopting contrastive learning to distinguish between "correctly focused" and "incorrectly focused" regions, the attention map can be made more in line with the prior knowledge of astronomers.

In addition, more physically meaningful aggregation strategies need to be designed, such as weighted fusion of features at different levels (e.g., core, petal, background) through structure-aware mechanisms, or exploration of explainability-driven aggregation functions (e.g., spatial-weighted pooling), to make the aggregation process itself interpretable.

Finally, standardized post-processing and interactive visualization tools should be developed to assist researchers in visually comparing the changes in attention maps under different settings, thereby better understanding model behavior and selecting appropriate configurations.

3.2.2 Improving the Efficiency of Multi-Branch Networks Through Lightweight Architecture and Automated Optimization

Firstly, in terms of model architecture, parameter sharing and a single backbone design can be adopted to replace the current dual-branch independent structure. For instance, a shared backbone network can be used to extract features, and two branches can be derived from different network layers: deep features are used to generate global context representations (replacing the Primary Branch), while shallow or mid-level features are upsampled and combined with attention maps to crop the regions of interest (replacing the Auxiliary Branch). Further, dynamic computational paths can be explored, where the activation of the auxiliary branch is adaptively determined based on the complexity of the input image. For simple samples, only the main branch is used for inference, thereby enhancing the average inference efficiency.

Secondly, to address the computational overhead of attention and Gaussian mask generation, a differentiable and lightweight alternative mechanism can be introduced. For instance, a learnable Spatial Transformer Network can be used to replace the manually designed Gaussian mask, allowing the network to automatically learn how to warp or crop the image to focus on the key regions.

Finally, to enhance the stability of training, a phased training strategy can be adopted. First, train the main branch and the attention module independently until they are stable, and then introduce the auxiliary branch for joint fine-tuning. Additionally, the introduction of gradient clipping, more precise loss weight scheduling, and normalization techniques can effectively alleviate the gradient instability issues caused by the joint training of multiple components.

4. Conclusion

This paper provides a comprehensive review of attention-based CNN methods for galaxy classification. This paper systematically surveys three prominent approaches—attention-gating, multi-branch attention networks, and dynamic multiscale attention networks. Highlighting their architectures, advantages, and performance. The analysis reveals that current methods still face challenges such as sensitivity to hyperparameters, high computational cost, and structural complexity. Future work should focus on developing more robust and interpretable attention mechanisms, lightweight architectures, and adaptive training strategies. Despite these limitations, attention-based CNNs represent a promising direction for achieving highly accurate and efficient automated galaxy classification.

References

[1] Conselice CJ, et al. The Evolution of Galaxy Number Density at z < 8. The Astrophysical Journal. 2016;830(2):83.

[2] Zwicky F, et al. Catalogue of Galaxies and of Clusters of Galaxies (Vol. I-VI). California Institute of Technology, Pasadena; 1961–1968.

[3] Storrie-Lombardi MC, et al. Morphological Classification of Galaxies by Artificial Neural Networks. Monthly Notices of the Royal Astronomical Society. 1992;256(3):427-436.

[4] Dieleman S, et al. Rotation-invariant Convolutional Neural Networks for Galaxy Morphology Prediction. Monthly Notices of the Royal Astronomical Society. 2015;450(2):1441-1459.

[5] Dai J, Tong J. Galaxy Morphology Classification Based on

ISSN 2959-6157

Deep Residual Network. Progress in Astronomy. 2022;41(2):45-58.

[6] Jiang J, Schmidt K, Valiante E, Pakmor R, et al. Galaxy Morphology Classification Based on Convolutional Vision Transformer. Astronomy & Astrophysics. 2024;683:A102.

[7] Bowles M, Scaife AMM, Porter F, et al. Attention-gating for Improved Radio Galaxy Classification. Monthly Notices of the Royal Astronomical Society. 2021;501(3):4579–4595.

[8] Mnih V, Heess N, Graves A, et al. Recurrent Models of Visual Attention. Advances in Neural Information Processing Systems. 2014;27.

[9] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997;9(8):1735–1780.

[10] Jetley S, Lord NA, Lee N, et al. Learn to Pay Attention. arXiv preprint arXiv:1804.02391. 2018.

[11] Tang H, Scaife AMM, Leahy JP. Classifying Radio Galaxies with Deep Convolutional Neural Networks. Monthly Notices of the Royal Astronomical Society. 2019;488(3):3358–3376.

[12] Porter F. MiraBest Batched Dataset[DS]. Zenodo; 2020. DOI: 10.5281/zenodo.4288837.

[13] Zhang Y, Liang G, Su Y, et al. Multi-Branch Attention Networks for Classifying Galaxy Clusters. arXiv preprint. 2021.

[14] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

[15] Jetley S, Lord NA, Lee N, et al. Learn to Pay Attention. arXiv preprint arXiv:1804.02391. 2018.

[16] Bellemare MG, Danihelka I, Dabney W, et al. The Cramér Distance as a Solution to Biased Wasserstein Gradients. arXiv preprint arXiv:1705.10743. 2017.

[17] Nelson D, Pillepich A, Springel V, et al. The IllustrisTNG Simulations: Public Data Release. Computational Astrophysics and Cosmology. 2019;6(1):1-24.

[18] Ma B, Qiu B, Luo A, et al. Galaxy Morphological Classification Using Dynamic Multiscale Attention Network. Monthly Notices of the Royal Astronomical Society. 2025;541:1928–1939.

[19] Radford A, Kim JW, Hallacy C, et al. Learning Transferable Visual Models from Natural Language Supervision. International Conference on Machine Learning. PMLR; 2021:8748–8763.

[20] Walmsley M, Lintott C, Géron T, et al. Galaxy Zoo DECaLS: Detailed Morphological Classifications from Volunteers and Deep Learning for 314,000 Galaxies. Monthly Notices of the Royal Astronomical Society. 2022;509(3):3966–3988.